Supplementary material for "SMOG@ctbp: Simplified deployment of structure-based models in GROMACS"

May 13, 2010

Contents

1 Supplementary Methods

1.1 Scalability of structure-based simulations using GROMACS

This is a description of the methods behind Figure 1 in the main text. To test the scalability of this forcefield over many processors, simulations were performed on 2, 4, 8, 16, 32, 64 and 128 cores on the Encanto Supercomputer (1792 nodes with dual 3.0 Ghz Quad core Intel Xeon Processors each and Infiniband interconnect). Each simulation was 10,000 timesteps in duration. The dynamic load balancing feature of GROMACS was used. For each number of nodes considered, 5 simulations were performed for each possible combination of domain decomposition grid settings. For example, when using 8 cores, simulations were performed with domain decomposition values of (2,2,2), (1,2,4), (1,4,2), (2,1,4), (2,4,1), (4,1,2), (4,2,1), (1,1,8), (1,8,1) and (8,1,1). Main text Figure 1 shows the highest performance settings averaged over 5 independent simulations.

We simulated a recent structure of the ribosome (PDB entries: 2WDG, 2WDH) to test the scalability of the forcefield to large systems. Prior to simulation, we removed the tRNA molecules (to avoid difficulties with aminoacylated-tRNA and modified nucleic acids) and magnesium ions from the PDB structure. In total, there were 142,196 atoms in each simulation. Since the removal of tRNA results in a structurally heterogeneous system (*i.e.* significant empty space), which can have adverse affects on the dynamic load balancing features of GROMACS, this study represents a lower bound on the potential scalability of structure-based models in GROMACS.

1.2 All-atom SBM scaling

V_{AA} has two free parameters, $R_{\mathrm{C/D}}$ and $R_{\mathrm{BB/SC}}$ which determine ϵ_{C}, ϵ_{BB}, and ϵ_{SC}. $R_{\mathrm{C/D}}$ is the ratio of total contact energy to total flexible dihedral energy while $R_{\mathrm{BB/SC}}$ gives the ratio of a single backbone dihedral to a single side chain dihedral. A side chain dihedral is any dihedral containing a non-backbone atom. The total stabilizing binding energy is normalized to the total number of atoms, $E_{\mathrm{S}} = N_{\mathrm{C}}\epsilon_{\mathrm{C}} + N_{\mathrm{BB}}\epsilon_{\mathrm{BB}} + N_{\mathrm{SC}}\epsilon_{\mathrm{SC}} = N_{\mathrm{A}}$, where N_{A} is the number of atoms, N_{C} is the number of contacts, N_{BB} is the number of backbone dihedrals, and N_{SC} is the number of side chain dihedrals. From $R_{\mathrm{C/D}}$, $R_{\mathrm{BB/SC}}$, N_{A}, N_{C}, N_{BB}, and N_{SC}, one can calculate ϵ_{C}, ϵ_{BB}, and ϵ_{SC}.

$$R_{\mathrm{C/D}} = \frac{N_{\mathrm{C}}\epsilon_{\mathrm{C}}}{N_{\mathrm{BB}}\epsilon_{\mathrm{BB}} + N_{\mathrm{SC}}\epsilon_{\mathrm{SC}}} \tag{1}$$

$$R_{\mathrm{BB/SC}} = \frac{\epsilon_{\mathrm{BB}}}{\epsilon_{\mathrm{SC}}} \tag{2}$$

$$\epsilon_{\mathrm{C}} = \frac{N_{\mathrm{A}}}{N_{\mathrm{C}}}\left(\frac{R_{\mathrm{C/D}}}{1 + R_{\mathrm{C/D}}}\right) \tag{3}$$

$$\epsilon_{\mathrm{BB}} = \frac{N_{\mathrm{A}}}{1+R_{\mathrm{C/D}}}\left(\frac{1}{N_{\mathrm{BB}}+N_{\mathrm{SC}}/R_{\mathrm{BB/SC}}}\right) \tag{4}$$

$$\epsilon_{\mathrm{SC}} = \frac{N_{\mathrm{A}}}{1+R_{\mathrm{C/D}}}\left(\frac{1}{N_{\mathrm{SC}}+N_{\mathrm{BB}}R_{\mathrm{BB/SC}}}\right) \tag{5}$$

2 Helpful Information

2.1 PDB file

Basically, follow standard PDB formatting. See the website for a sample PDB file, so you can see the format expected by the webtool. Your PDB file MUST conform to the following standards.

- No hidden characters. They can lead to unpredictable results. To avoid accidentally inserting them use a text editor such as `vi` or `emacs`.
- If your file does not work with the webtool, only include lines that start with `ATOM` (to specify each atom), `TER` (to indicate a break between 2 chains) and `END` at the end of the file.
- Chain identifiers are not used. If you have multiple chains, insert `TER` (left justified) between chains. The webtool will internally index the chains sequentially, starting with 1.
- Terminal oxygens (in proteins) are called `OXT` and `O` (not `O1` and `O2`).
- The file is not read past an `END` statement (ALL CAPS, left justified). If atoms appear after an `END`, these atoms will not be included.

Recognized residues include:

- Protein residues: All 20 amino acids (3 letter codes).
- RNA residues: CYT or C, GUA or G, URA or U and ADE or A.
- DNA residues: DG, DC, DA, DT.
- Ligands: SAM (S-Adenosylmethionine), GNP (Gpp(NH)p), ATP, ADP, AMP

2.2 Contact maps

A contact map is a list of atom-atom pairs that are "in contact" in the native structure (the PDB structure). These pairs will interact via Lennard-Jones interactions with the energetic minimum at the distance found in the PDB structure. There are three supported ways of defining a contact map.

- **Shadow map**: A shadow map includes contacts that are within a cut-off distance, are separated in sequence by at least 3 residues, and do not have an atom in between them. A full description can be found on the web server and in a forthcoming publication. If you select this option, a shadow map will be generated and used in the Hamiltonian, with default values.
- **Cut-off map**: This will generate a list of contacts as determined by your specified distances and sequence differences. Recommended values are the defaults. This option is NOT enabled for C_α model.

- **Upload file** (Upload your own contact map): The contact file requires the following format: Each line identifies a single contact. Each line has 4 fields (chain i, atom number i , chain j, atom number j). For example, to include a contact between the atom 10 (PDB numbering) of the first chain (internally indexed as 1) and atom 20 of the third chain, the line would read: `1 10 3 20`

Blank lines, even at the end of the file can cause trouble. If you are using the C_α model, use residue numbers in the contact map, not atom numbers.

2.3 Stacking interactions in RNA

When using the 4Å cut-off definition, all atom-atom pairs that are within 4Å in the native structure are assigned as contacts. According to this definition of a contact, stacked bases (bases adjacent in sequence) have about 5 times the number of contacts that hydrogen bonded base pairs have. To account for this, the contacts between stacked bases are rescaled by 1/3. Stacking interactions are chemically different from hydrogen bonding. When using a cut-off distance criterion for contacts, there is no reason that each stacking contact should be the same strength as a base pairing contact. The rescaling is only intended to give a more reasonable distribution of energy.

2.4 Graining

There are currently two levels of graining available, all-atom and C_α.

- **All-atom**: The all-atom model has an explicit bead for each heavy atom. If you have hydrogens in your PDB file, they SHOULD be ignored by this program. But, if you have given hydrogens non-standard names, then the program may complain. All beads are the same size. Since the geometry is explicitly represented, bonds, bond angles, and dihedrals have their traditional meanings. Contacts are defined between native atomic pairs. The exact choice of parameters is up to you. The default values on this page are suggested values. Always double check your choice of parameters. A complete description of the all-atom model can be found here:
  - **For proteins**: Whitford, P.C., Noel, J.K., Gosavi, S., Schug, A., Sanbonmatsu, K.Y. & Onuchic, J.N. (2009). An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. *Proteins* **75**, 430-41.
  - **For RNA/DNA**: Whitford, P.C., Schug, A., Saunders, J., Hennelly, S.P., Onuchic, J.N., Sanbonmatsu, K.Y. (2009). Nonlocal Helix Formation Is Key to Understanding S-Adenosylmethionine-1 Riboswitch Function. *Biophysical Journal.* **96** L7-9.
- **C_α**: The C_α model is defined only for proteins. If RNA or DNA is present the all-atom model must be used. The C_α model has a single bead-per-residue centered at the location of the C_α. Bonds, angles and backbone

dihedrals are between two, three and four beads, respectively. Backbone dihedrals and contacts are equally weighted and contacts are defined between native residue contacts. A complete description of the C_α model can be found here:

– Clementi C., Nymeyer H. & Onuchic J.N. (2000) Topological and energetic factors: What determines the structural details of the transition state ensemble and ”en-route” intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **298** 937-953.

2.5 Contact to dihedral energy ratio in all-atom model

Also see Section 1.2. This is the ratio of the total stabilizing energy in ALL contacts and the total stabilizing energy in ALL flexible dihedrals (*i.e.* not ring, improper, or fixed dihedral angles). This quantity is fully described elsewhere. The sum of the strengths of all contacts and all dihedrals is then normalized to the number of atoms in the system.

- **Protein backbone to sidechain dihedral ratio**: In this model, all protein backbone dihedral angles are given identical energetic weighting. All protein side chain dihedrals are also given identical energetic weighting. This quantity sets the ratio of the strength of a single protein backbone dihedral to a single protein side chain dihedral.

- **RNA/DNA backbone to sidechain dihedral ratio**: In this model, all RNA/DNA backbone dihedral angles are given identical energetic weighting. All RNA/DNA side chain dihedrals are also given identical energetic weighting. This quantity sets the ratio of the strength of a single RNA/DNA side chain dihedral and a single RNA/DNA backbone dihedral.

2.6 Excluded volume

One important feature of the model is the size of the atoms. Two parameters, ϵ_{NC} and σ_{NC}, determine the excluded distance between non-native pairs and the strength of the repulsive term. We used 0.01Å and 2.5Å in our initial papers, but we leave these parameters flexible for additional investigation.

2.7 Folding Temperatures

The folding temperature T_{F} can be defined in many ways: the peak temperature in C_{V}, when the free energy of the folded and unfolded states are equal, when the integrated probability of the unfolded basin equals that of the folded basin, *etc.* Since all these definitions give temperatures within 1% of each other for 2-state proteins with our structure-based models, it is unimportant which definition is chosen. We generally define T_{F} as the temperature when the free energy of the folded and unfolded states are equal. In the all-atom model, with the total stabilizing energy normalized to the total number of atoms, the folding temperature of globular proteins is generally ~ 1 in reduced units. The C_α model also generally has folding temperatures near 1 in reduced units.

2.8 Changes to the published parameter values

The default values are the values used in the initial protein and RNA papers, with the following modifications:

- The harmonic dihedral angle constant that maintains planarity of rings has been increased from 10 to 40. This makes the rings more rigid.

3 GROMACS

3.1 Reduced Units

The structure-based model is run in reduced units. This means that the length scale, time scale, mass scale, and energy scale are all 1. The natural GROMACS units are length scale nm, time scale ps, mass scale amu, and energy scale kJ/mol. We convert the PDB (length scale Å) into nanometers but the mass scale, time scale, and energy scale are "free." You could, in theory, determine an overall energy scale and mass scale from the structure and dynamics, and then infer a time scale, but this should be performed with care. This estimate of time does not take into account the effects of solvent friction and possible energetic roughness that may slow down the dynamics of a real system. There is no standard method of computing "real" times from structure-based simulations. For the interested reader, we suggest Kouza et al. http://pubs.acs.org/doi/abs/10.1021/jp053770b. Essentially, be careful. At least you can be sure that the "real" time unit is longer than the picosecond time scale assumed in GROMACS.

One must take care with the temperature units in GROMACS. In reduced units $k_\mathrm{B} = 0.00831451$. The temperature sets an energy scale $k_\mathrm{B}T$. So to have a reduced temperature of 1, you must use a GROMACS temperature of $1/0.00831451 = 120.2717$. This means to be near folding temperature, the value of temperature in the `.mdp` file will usually be around 120.

3.2 Brief GROMACS Tutorial

Since there are already many GROMACS tutorials available online, this tutorial only covers the absolute basics of GROMACS, plus special instructions (of which there are very few) for structure-based simulations.

After uploading a PDB file and pressing the "Submit" button, the web server will either return a link to the completed output or return an error message describing any formatting inconsistencies. The completed output is a tarball containing:

1. GROMACS coordinate file: the initial structure corresponding to the provided PDB structure; shifted such that the box starts at the origin (`.gro`).
2. GROMACS topology file: describes all the atomic interactions in the SBM Hamiltonian (`.top`).
3. GROMACS index file: convenient for manipulating structures with multiple chains (`.ndx`).
4. Native contact map: if Shadow is selected (`.contact`).
5. Web server output: contains any non-fatal warnings and messages (`.output`).

In order to run a simulation in GROMACS, you need three files: a parameters file (`.mdp`), a topology file (`.top`), and a coordinate file (`.gro`). The `.mdp` file tells GROMACS what settings you would like to use for the simulation, *e.g.* the temperature, the time step, and the temperature coupling constant. A sample `.mdp` file for structure-based models is available on the web server. The topology file gives GROMACS all of the specifics of your Hamiltonian. In our case, the `.top` file tells GROMACS about the interactions that define a structure-based model. The `.gro` file tells GROMACS the initial coordinates of the atoms. It also states the (x,y,z) size of the simulation box as the last line.

Once you have these three files, there are only two GROMACS commands necessary to run your simulation.

First, produce a portable `xdr` file (in this case, `run.tpr`) that describes your simulation. This file is platform independent and contains all parameters for your simulation. This allows you to produce a `.tpr` file on any machine, and then move it to another machine and run your simulation. The `xdr` file is produced by `grompp` (binaries are located in the compiled GROMACS distribution, usually located in `${GROMACSroot}/bin/`):

- `grompp -f mdp_file.mdp -c gro_file.gro -p top_file.top -o run.tpr`

Run an all-atom simulation by calling the molecular dynamics module of GROMACS and tell it to read `run.tpr`:

- `mdrun -s run.tpr`

Or, run a C_α simulation:

- `mdrun -s run.tpr -table table_file.xvg -tablep table_file.xvg`

If you are running a C_α model, where a 10-12 potential is used instead of a 6-12 potential, then you MUST: 1) provide the 10-12 lookup table, 2) indicate that you want to use the tables by modifying your `.mdp` file and 3) issue `mdrun` with the flags noted above. A sample C_α `.mdp` file can be found on the web server. There also is a Perl script to make Van der Waals tables for GROMACS version 4.0.X. WARNING: Using a table for v4 with v3 will cause serious problems.

To run a simulation in parallel, use the command:

- mpirun -np ${NUMBER_OF_NODES} mdrun -s run.tpr -noddcheck

The second flag `-noddcheck` is necessary with this model because our model can confuse the dynamic load balancing utility of GROMACS 4. Additionally, you need to determine which version of MPI is best for you. This flag merely tells GROMACS not to worry. Running a simulation in parallel is highly dependent on your machine's configuration. You can consult the GROMACS help page for more information on the details of parallel simulations.

4 Frequently Asked Questions

4.1 Do I need a specific version of GROMACS?

We highly recommend using GROMACS 4.0.X as there are significant improvements over earlier versions. The web tool should provide valid topology and coordinate files for GROMACS 3.3 but we will only offer support for GROMACS 4.0.X.

4.2 Why do I need space between the system and the boundaries?

This software will recenter the system in a box with space between the system and the boundary. The box starts at the origin. This is important when using periodic boundary conditions (which is required for grid neighbor searching in GROMACS). Make sure your box is large enough for the dynamics of interest. For example, if you are going to look at folding/unfolding, make sure you have a large buffer so that the molecule will not extend out of the box and interact with its own image. Pay particular attention to this settings if you are using a C_α model, where the non-bonded cut-off distance is large. Additionally, if you are running parallel simulations, the load balancing has trouble (highly reduced performance) if there is a lot of empty space, so it is a good idea to use the smallest box size that is practical.

4.3 When running `grompp`, why does GROMACS call RNA residues "DNA" and DNA residues "other"?

This does not mean anything is wrong with the simulation. Our software uses PDB naming for our residues. i.e. RNA is called U, A, C and G and DNA is called DT, DC, DG and DA. But, GROMACS expects DNA to be called U, T, C and G. This can also happen when porting AMBER nucleic acid forcefields to GROMACS.

4.4 The webtool says I am missing bonds, angles, or dihedrals. Should I be concerned?

This indicates that your structure (A) is missing atoms or (B) has non-standard atom names. Since the webtool expects certain atoms (with specific bonds, angles and dihedrals) to be present, if it doesn't see those atoms then it can't include the associated terms. If atoms are missing then the webtool will produce a file where those atoms, and all associated energetic terms, are excluded. The files provided will still run in GROMACS. If you believe you have a complete structure in the PDB file and you get these messages, check the output coordinate (`.gro`) file and see what atoms are missing. Make sure the atom name field is formatted correctly. The exact location of the atom name is important (*i.e.* "` CD1`" is not the same as "`CD1 `").

4.5 Why do the same messages appear twice in the output file?

This happens when you are using a shadow map. It happens because we have to process the PDB file twice when using the Shadow map. The second set of messages should reflect the options you selected. The first set of messages are set to default values, and have no bearing on the final `.gro` and `.top` files.

4.6 Why do the messages in the output file indicate options different from what I chose?

As indicated above, this is due to selecting a Shadow map. The only settings that affect your simulations will be the second set of messages.

4.7 I am using the C_α model and I see unexplanable spikes in the energy. What is going on?

This is a C_α-related issue. In the C_α model, bond angles are defined by adjacent C_α atoms and not by real bond angles. This means the angles can be rather large. In GROMACS (and other MD packages) there is a numerical instability when a bond angle is 180 degrees because the force is undefined for such values. This issue appears mostly in GROMACS v3. This issue was reduced in GROMACS v4. It is rare that this happens, but if it does in your system, you can increase the strength of all bond angles. This will reduce the chances of reaching 180 degrees and the spikes should go away. If you do this, make sure your time step is not too large for the new strength of the angles.

4.8 Can I download the source code?

We are currently working on a nicer, more flexible, version of the code that drives this webtool. We will post it on this site when it is ready for distribution. In the meantime, the current code is available upon request.